**TMM026- Introductory Data Science**

*(12.3.2021)*

**Typical Range:** 4-5 Semester Hours

**Pre-Requisite Recommendations:** No pre-requisite will be required

**Rationale**:

Introductory Data Science (IDS) is an emerging field that uses methods from statistics, mathematics, and computer science to find and communicate meaning in data. Due to the rapidly increasing role of data in commerce and science, data science has gained wide attention in a relatively short amount of time in the private sector, government, and academia. In the private sector, job advertisements for data scientists have proliferated.

There are at least three program areas centered on IDS that could satisfy their cohorts.

i. Associate degree programs for students who intend to transfer to a four-year institution.
ii. Associate degree programs for students aiming to go directly into the workforce.
iii. Credit bearing certificate programs.

These three different types of programs share many aspects but differ in the emphasis placed on each of these program outcomes. All have IDS as a focal point of study.

The goal of this course is not to transform each student into a data scientist, but to give the student a sense of data literacy. That is, the ability to collect, analyze and derive meaningful information from data. This course does not require prior mathematical, statistical, or programming skills.

**Topic Areas**:

1. **Curation of Data** – Data management and curation is a first step in IDS. This includes acquiring data from diverse formats and structures, cleaning data to prepare for data analysis and maintaining and sharing data files in a version control system.
2. **Enhanced Data Visualization** – Statistics and IDS often tell a story of data through informative pictures. Enhanced data visualizations go beyond the common graphs in introductory statistics to describe, explore and communicate insights from data.
3. **Statistical Models, Estimation, and Prediction** – Determining a functional relationship between numerical data and numerical/categorical data is at the center of statistical modeling in IDS. The focus is on using statistical models to describe relationships between variables, discerning between modeling for predictions and modeling for inference, and fitting, evaluating, and interpreting statistical models. Students familiar with statistical inference from introductory statistics are unavoidably sheltered from the mathematical rigor behind each hypothesis test. Simulation based probability and inference provides straightforward methods to make decisions with data without working through the mathematical details that underlie traditional statistical inference. The quality of a prediction model in terms of being able to forecast an expected outcome is measured through a loss function.

4. **Applications of Data Science** – Machine learning and statistical learning. Machine learning is the process of developing computer algorithms to search for and recognize patterns in data. Statistical learning additionally uses the expertise of a human analyst to craft appropriate models. There are two branches in machine/statistical learning: supervised learning and unsupervised learning.
5. **Consumer of Data Science** – The importance of ethical data science practice is critical to the validity of results in any IDS course. This includes problem-solving and the use of 'big data' which can accentuate cultural biases and differences and discussion of issues involving privacy, data security and societal impact.

To qualify for TMM026 (Introductory Data Science), a course must achieve all the following essential learning outcomes listed in this document (marked with an asterisk). These make up the bulk of an Introductory Data Science course. Courses that contain only the essential learning outcomes are acceptable from the TMM026 review and approval standpoint. It is up to individual institutions to determine further adaptation of additional course learning outcomes of their choice to support their students' needs. As time permits, institutions are welcome to incorporate non-essential learning outcomes (those not marked with an asterisk) or additional institutional specific learning outcomes. In addition, individual institutions will determine their own level of student engagement and manner of implementation. These guidelines simply seek to foster thinking in this direction.

1. **Curation of Data-** Data curation is a way of managing data that makes it useful for users who are interested in that data's analysis. Data curation is an active and continuous process of the management of data thought its time of usefulness. Successful data science students can curate large multivariate data sets from various sources to ready them for analysis.

   The successful Introductory Data Science student can:

   **1.1. Types and Sources of Data**- Distinguish between types and sources of data. *

   **Sample Tasks:**

   - Differentiate between categorical and quantitative data.
   - Classify data as nominal level, ordinal level or ratio level and discuss its characteristics and limitations.
   - Discern between a set of data and a source of data.

   **1.2. Collecting Data-** Acquire raw data, the first step in any comprehensive analysis of data and its associated topic. These data sources can come from data bases, flat files, web services or other sources such as designed surveys and RSS feeds. *****

   **Sample Tasks:**

   - Distinguish between different sources of data such as relational database, automated data collection, and online surveys.
   - Discern between structured data sources, sources that are searchable such as relational databases, and unstructured data sources, sources that are not searchable such as social media and text messages.

- Collect data from open or public data sources such as data.gov, IPUMS, Kaggle, Quandl, The World Bank, US Census Bureau, NASA, Amazon Web Services or Google Cloud Platform.
- Convert a file from its present format into a format that is prepared for analysis.

**1.3. Organizing and Standardizing Data-** Ensure the clarity, completeness, and stability of the data through the organization of that data, which is a significant aspect of data curation. *****

**Sample Tasks:**

- Combine data sets for comprehensive use and analysis.
- Wrangle data by labeling variables and sorting, filtering, and arranging data so that it is prepared to answer a given data analytical question.
- Scaling data using Simple Feature Scaling and Min-Max Scaling.
- Normalizing data using Z-score normalization.

**1.4. Cleaning Data-** Engage in the processes of identifying incorrect, incomplete, inaccurate, irrelevant, or missing data and then modifying, replacing, or deleting that information as needed. Data cleaning is considered a fundamental step in introductory data science. *****

**Sample Tasks**:

- Explain why some set of data has missing values and how to account for the missing data.
- Detect outliers using graphical methods such as boxplots
- Classify outliers as errors, missing values, or unusual values.
- Clean data as necessary and eliminate variables deemed as irrelevant.
- Use a package such as tidyr in R to clean raw data sets.
- Enhance raw data as necessary by converting time zones, making currency conversions, calculating values or time units.
- Address outliers in a set of data using mathematical techniques such as log transformations or interpolation, or by deductive correction or deterministic imputation for missing values.

**2. Enhanced Data Visualization-** A successful data science student will choose an appropriate presentation model to fit a given topic, and then present results using a variety of charting and graphing techniques.

The successful Introductory Data Science student can:

**2.1. Traditional Plots-** Classify and summarize data using traditional plots**.** *****

**Sample Tasks:**

- Before undertaking any activity, students must answer the following questions:
  - What are you trying to show?
  - Does a trendline, heat map, time series, etc. make sense in this case?
  - Why will one type of graph work better than another?

- Emphasize that the data and analysis must tell the story, and the charts are a helpful tool to tell the story.
- State with clarity what you are trying to say.

**2.2. Single vs. Two or More Variables-** Select appropriate charting techniques based on the type of data and the number of variables they intend to present. *****

**Sample Tasks:**

- Discuss differences between numerical and categorical data.
  - What types of charts are appropriate for numerical data?
  - What types of charts are appropriate for categorical data?
- Effects of outliers
  - Explain the difference between an explanatory variable and a response variable.
  - Generate a general hypothesis about the relationship between two variables.
  - Construct a scatter plot using a large data set containing 1000+ points.
  - Confirm that a trendline is appropriate for the data.
  - Build a linear model using the trendline.
  - Examine the outliers and decide if they should be included in the model.
  - If appropriate, remove the outliers, adjust model, plot a new trendline, and build a new model.
  - Compare the two plots, trendlines, and models.
  - Summarize the effects of the outliers on the response variable.

**2.3. Dynamic Visualization-** Compare traditional and dynamic graphing techniques and give reasons for and justify why a dynamic plot may be the appropriate choice (or is the appropriate choice for a specific data set). *****

**Sample Tasks:**

- Decide the appropriate type of graph or illustrations based on the data set
  - Discuss the similarities and difference between a traditional boxplot and violin plot.
    Consider advantages and disadvantages of using a map plot versus a heat map
- Tell the story using graphs
  - The goal of the activity is for the student to be able to create a presentation and produce a series of graphs using a large data set.
  - Students brainstorm ideas about what they would like to study.
  - Perform preliminary research and search for available data sets.
  - Generate an informal hypothesis based on their preliminary findings.
  - Gather and clean data.
  - Do trial charting and choose appropriate graphs.
  - Examine outliers and reclean the data if necessary and appropriate.
  - Generate a formal hypothesis based upon what they've been seeing in the data set.

- Explain why the hypothesis fits their "story."
- Create a series of charts that verifies your hypothesis
- Present your findings to tell the story.
- Include the caveat that generation of a formal hypothesis and confirmation of the hypothesis on the same set of data does not settle the question.

**2.4. Distribution Maps-** Identify common distribution models and discern what types of data fit certain models. *****

**Sample Tasks:**

- Student will be able to create distribution maps of quantitative variables
  - Biological data that fit the normal distribution such as blood pressure, height, and weight
  - Data that does not fit the normal distribution such binomial or Bernoulli distributions
  - Students analyze wait times and build an appropriate distribution map based on their observations.
  - Student will investigate heat maps and choose to highlight different variables.
- Voting patterns:
  - Student will find a large data set of voting information for a certain state or county.
  - Students create maps to display the voting information.
  - Students make a hypothesis about voting patterns in a certain geographic area.
  - Student will analyze the different areas and a manipulate the fineness of measurement to highlight various areas.
  - Based on the observations, the students will verify the hypothesis.
  - Students will describe a future data set that would allow them to confirm their hypothesis.
- Plotting a data set with large variation for a market study
  - Find a data set of ages for customers that visit McDonalds for breakfast
  - Plot the entire data set using an appropriate chart
  - Look for a pattern (we hope that it does not exist)
  - Generate a hypothesis that could pertain to breakfast promotion
  - Break the data into subsets
    - Age range
    - Time of morning
    - Day of the week
  - Replot using the subsets
  - Look for a pattern to verify your hypothesis
  - Summarize the process
  - Present as a marketing pitch

**2.5. Time Series Plots-** Develop an analytic model and trendline, and then predict the last n-tile of data in order evaluate the effectiveness of their model. *****

**Sample Tasks:**

- Avoid extrapolation
    - Build a model using a weather data set from 1900 to present.
    - Create a scatter plot using rain data and add a trendline.
    - Predict rainfall 1 week and 1 month into future.
    - Verify the accuracy of your model using the actual rainfall.
    - Add the updated data to your chart, and check to see if it falls within a standard deviation of the prediction.
    - Discuss why the prediction for 1 week was more accurate than for 1 month.

**2.6. Misleading Graphs-** Locate data visualizations and deconstruct the graph in order to evaluate the effectiveness of the visualization. *****

**Sample Tasks:**

- Find graphics in the "wild" and bring them into class to critique.
    - Example of bad graphs could be the following:
        - Non-zero axis
        - Zero axis
        - Pie charts versus bar charts versus histograms
        - Inappropriate pictograms
    - The goal of the project is for students to find intentionally misleading graphs.
    - Students will critique each example of a bad graph and explain why it is inappropriate in the given situation.
    - Students will propose a more appropriate presentation for the same data and explain why that presentation is the better choice.
    - Students will proceed to create an example using the appropriate presentation.
    - In the end the students present the inappropriate and appropriate graphs to the class for side-by-side comparison.
- Choosing to highlight proper information.
    - Student will choose variables they want to highlight from a larger set of variables.
    - Student will need to decide what they want to show.
        - Cases
        - Hospitalizations
        - Deaths
    - Student will explain the importance of one set of variables versus another.
        - Student will experiment with different color-coding schemes to highlight important variables, and decide between the following:
            - Standard pallet
            - Custom pallet
        - Students will justify their choice.
        - Student will discuss disability issues such as color blindness.

**3. Statistical Models, Estimation, and Prediction** - Students will develop the ability to write generative models for data and to use them with simulation-based approaches. They will use these models to develop an understanding of variation, to approximate probabilities, and to

diagnose and repair shortcomings of a statistical model. Using the sample data from a statistical model, students will be able to estimate the parameters of the model and to make predictions for new data related to the statistical model. In addition, students will formulate statistical inference as a decision problem, will understand the concept of a loss function, and will use simulation methods to select an estimator that has small expected loss under the decision problem.

The successful Introductory Data Science student can:

**3.1. Statistical Model -** Write and implement generative models for situations ranging from simple one-sample problems to more complex settings. *

**Sample Tasks:**

- Generate data from standard probability distributions such as the binomial distribution, the Poisson distribution, the normal distribution, and the gamma distribution.
- Draw a graphic (histogram / density estimate) to describe a sample of data.
- Write a statistical model that includes both structured components and stochastic components such as a simple linear regression model with normal errors, a multiple linear regression model, or a logistic regression model.
- Generate data from a complex statistical model.
- Write a statistical model that includes more than one stochastic component such as a mixed model for linear regression.

**3.2. Estimator and Sampling Distribution** – Understand how to estimate the parameters of a model by summarization of a data set, to use simulation methods to evaluate rival estimators, and to describe the bias and variance of an estimator in the sampling distributions of estimators. *

**Sample Tasks:**

- Identify the parameters of a statistical model to be estimated.
- Compute one-sample summaries (estimates) of the center of a distribution – mean, median, and trimmed mean.
- Simulate samples of data from a generative model, summarize the sample with an estimator, and examine the estimator's sampling distribution (histogram / density estimate), its center and its spread.
- Use the center and spread of the sampling distribution to describe the accuracy of an estimator in terms of bias and variance.
- Simulate data to explore and evaluate sampling distributions of estimators in complex statistical models.

**3.3**. **Simulation-Based Estimation and Prediction** - Use simulation methods to understand the implications of statistical models. *

**Sample Tasks:**

- Estimate the probability of an event for a simple probability model from a simulation.
- Assess the accuracy of an estimated probability.
- Estimate the probability of an event in a complex statistical model from a simulation.
- Estimate the conditional probability of one event given another event for models with a single component of variation and for models with more than one component of variation.
- Estimate implications (features) of a statistical model. These features might be a percentile of a distribution for a specified value of the predictor, the regression coefficient itself, or the odds for logistic regression.
- Make a prediction for new data from the model.

**3.4. Loss Function and Decision -** Formulate statistical inference as a decision problem, specify a generative statistical model, a target of inference, and a loss function, and then select an estimator by minimization of average loss over simulated data sets.

**Sample tasks:**

- Choose among mean, median, and trimmed mean as an estimator in a simple model, when the target is the center of the distribution and the loss function is squared error loss. (Note that the choice will depend on the generative model).
- Understand 0-1 loss for a two-choice problem.
- Demonstrate that the loss function determines the target of estimation (e.g., estimation of a quantile can be accomplished by minimization of an asymmetric absolute error loss function).
- Make use of squared error predictive loss in a simple linear regression setting.
- Design a loss function for a problem of the student's choosing.

**3.5. Model Diagnostics -** Diagnose and repair shortcomings of a statistical model. Contrast the implications of a statistical model with subject matter knowledge to identify its shortcomings. Improve the model by introducing new predictors, by altering the probability distribution used in the model, or by creating a more complex probability structure.

**Sample tasks:**

- Contrast the implications of a statistical model with subject matter knowledge to identify flaws in the model.
- Incorporate a new predictor into the model to reflect subject matter knowledge.
- Replace one probability distribution with another to modify the model.
- Introduce an additional component of variation to create dependence in observed data.

**3.6. Estimation (mathematical approaches) -** Become aware that analytic calculation can often replace simulation to provide a formal solution to the problems above.

**Sample tasks:**

- Describe a simple model for a game of chance such as the roll of a fair die.

- Describe probability as “area under the density curve” for continuous random variables.
- Compare the mathematical (analytical) approach and simulation-based approaches to estimation.

4. **Applications of Data Science-** Applications in the field of data science can vary by the topic at hand. Data science techniques can be applied to subject areas such as healthcare, business and commerce, education, transportation, sports and recreation, public policy, law enforcement, and social life. However, there are similar data analysis techniques that can be applied to all these disciplines**.**

The successful Introductory Data Science student can:

**4.1. Machine Learning-** Define machine learning and statistical learning, as well as differentiate between supervised and unsupervised learning. *

**Sample Tasks:**

- Identify, or give an example of, an unsupervised learning technique.
- Use a package such as caret in R to perform a machine learning algorithm.

**4.2. Supervised Learning -** Classify data using machine learning techniques, search for and define a function that describes how different measured variables are related to one another and utilize predictive techniques such as simple linear regression and multiple linear regression. *

**Sample Tasks:**

- Differentiate between supervised and unsupervised learning.
- Identify, or give an example of, an unsupervised learning technique.
- Identify, or give an example of, a supervised learning technique.
- Classify data using K-nearest neighbors.
- Classify discrete data using the Naive Bayes algorithm.
- Use simple linear regression analysis to predict the value of a response variable based on a given explanatory variable.
- Interpret the y-intercept and make inferences about the slope of a simple linear regression equation.
- Evaluate the assumptions of regression analysis and know what to do if the assumptions are violated.
- Interpret the correlation coefficient.
- Describe the purpose of multiple linear regression.
- Input variable information and data for multiple linear regression.
- Describe and discern the data assumptions required for multiple linear regression.
- Interpret scatterplots and probability plots concerning the data assumptions for multiple linear regression.

- Write a prediction equation and make predictions based on a multiple linear regression model.
- Use a command such as lm() in R to perform multiple linear regression.
- Use logistic regression to describe the relationship between an explanatory variable and a dichotomous response variable.
- Compare and contrast logistic regression and ordinary least squares regression.
- Fit a logistic model and use the model to estimate the odds from a single probability.
- Describe the statistical model of logistic regression with a single explanatory variable.
- Identify the estimates of the regression parameters and write the equation for a fitted model.
- For a given logistic model, compute and interpret the threshold value.
- Use a command such as glm() in R to perform logistic regression.

**4.3. Unsupervised Learning** Use algorithms to draw inferences from datasets consisting of input data without labeled responses. *

**Sample Tasks:**

- Identify data that is relevant to K-means clustering.
- Describe the basic steps of the K-means clustering algorithm.
- Interpret an elbow graph to determine the optimal number of clusters.
- List the advantages and disadvantages of K-means clustering.
- Use a command such as kmeans() in R to solve applications of K-means clustering.

**4.4. Sentiment Analysis-** Interpret and classify emotions within text data using rule-based or machine learning algorithms which focus on polarity (negative, neutral, or positive), feelings and emotions (happy, angry, sad, etc.) and intentions (interested or not interested).

**Sample Tasks:**

- Identify the various types of sentiment analysis.
- Identify, or give an example of, uses of sentiment analysis.
- Read text from a dataset and tokenize the data.
- Use a sentiment lexicon to analyze the sentiment for given text data.
- Visualize the sentiment of text data using scatterplots or boxplots.
- Use a package such as Sentiment Analysis in R to perform sentiment analysis for a given set of data.

**5. Consumer of Data Science-** Ethics can be thought of as the shared values which help humanity differentiate right from wrong. The access to big data, coupled with an increased awareness of information technology and data analysis have become an important aspect of the human experience. Consequently, data science has the ability to influence how decisions are made in business, life and health sciences, transportation, and civics. The successful data science student will be able to successfully differentiate between ethical and non-ethical decisions based on data.

The successful Introductory Data Science student can:

**5.1. Laws and Regulations** –Consider the local legislation, and identify the relevant laws, rules, and regulations pertaining to protection of personal data. *

**Sample Tasks:**

- Describe the European General Data Protection Regulation (GDPR) and explain how this regulation affects data science projects.
- Summarize Section 230 of the 1996 US Communication Decency Act and explain the consequences of how this act shields online publishers from liability of generated content.
- Compare and contrast the GDPR with the California Privacy Act of 2018 and similar legislation.

**5.2. Unfair Discrimination and Social Bias -** Discern bias from fairness in finance, medicine, and society in order to prevent incorrect or distorted conclusions. *

**Sample Tasks:**

- Identify data that reflects an unwarranted bias.
- Illustrate the reinforcement of human biases in computer models that make predictions in areas such as medical insurance, financial loans, or policing.
- Describe the effects of biased data as it relates to red-lining.

**5.3. Transparency of Methods and Analysis** – Identify clarity in methods of analysis of data and demonstrate how conclusions can be misleading. *

**Sample Tasks:**

- Differentiate between settings where an algorithm will lead to biased results and to unbiased results.
- Identify the traits that demonstrate that the analysis, findings, and conclusions made from data are reliable and reproducible.
- Explain the conclusions of the analysis of data in terms that are understandable to an appropriate audience.
- Identify misinterpretations in conclusions of data analysis.

**5.4. Sampling Bias** – Cite bias in its various forms. *

**Sample Tasks:**

- Identify data which is “cherry picked.” That is, selecting data that supports a particular position while ignoring relevant contradictory evidence.
- Characterize a sample of data influenced by nonresponse bias and explain that this bias is influenced by the salience of a study or its social appeal.

- Explain and pinpoint studies that could be affected by interviewer bias or from a voluntary response sample.